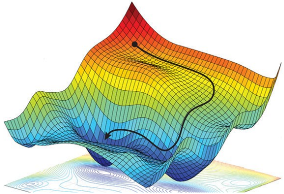
Introduction to Machine Learning

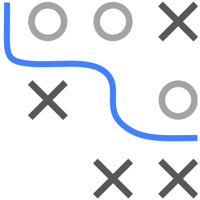
Risk Minimizers



Learning goals

- Know the concepts of the Bayes optimal model (also: risk minimizer, population minimizer)
- Bayes risk
- Consistent learners
- Bayes regret, estimation and approximation error
- Optimal constant model

EMPIRICAL RISK MINIMIZATION



Very often, in ML, we minimize the empirical risk

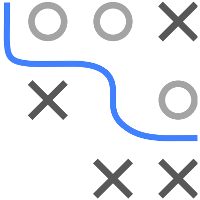
$$\mathcal{R}_{\text{emp}}(f) = \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right)$$

where

- each observation $\left(\mathbf{x}^{(i)}, y^{(i)}\right) \in \mathcal{X} \times \mathcal{Y}$ and $\mathcal{X}, \mathcal{Y}$ are the feature and target spaces, respectively,
- $f : \mathcal{X} \to \mathbb{R}^g, f \in \mathcal{H}$ is a model which maps a feature vector to a numerically encoded element of $\mathcal{Y}$ and $\mathcal{H}$ is the hypothesis space,
- $L : (\mathcal{Y} \times \mathbb{R}^g) \to \mathbb{R}$ is the loss function which measures the dissimilarity of the model prediction and the true target,
- and we assume that $\left(\mathbf{x}^{(i)}, y^{(i)}\right) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{xy}$ where $\mathbb{P}_{xy}$ is the distribution of the data generating process (DGP).

What is the theoretical justification for this procedure?

RISK MINIMIZER



Our goal is to minimize the risk

$$\mathcal{R}_L(f) := \mathbb{E}_{xy}[L(y, f(\mathbf{x}))] = \int L(y, f(\mathbf{x}))\, d\mathbb{P}_{xy}.$$

for a certain hypothesis $f(\mathbf{x}) \in \mathcal{H}$ and a loss $L(y, f(\mathbf{x}))$.

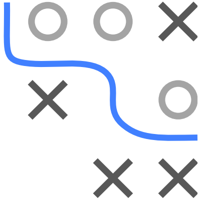
NB: As $\mathcal{R}_L$ depends on loss L, we sometimes make this explicit with a subscript if needed and omit it in other cases.

Let us assume we are in an "ideal world":

- The hypothesis space $\mathcal{H}$ is unrestricted. We can choose any $f : \mathcal{X} \to \mathbb{R}^g$.
- We also assume an ideal optimizer; the risk minimization can always be solved perfectly and efficiently.
- We know $\mathbb{P}_{xy}$.

How should f be chosen?

RISK MINIMIZER / 2



The f with minimal risk across all (measurable) functions is called the **risk minimizer**, **population minimizer** or **Bayes optimal model**.

$$\begin{aligned} f^* &= \arg\min_{f:\mathcal{X}\to\mathbb{R}^g} \mathcal{R}_L(f) = \arg\min_{f:\mathcal{X}\to\mathbb{R}^g} \mathbb{E}_{xy}\left[L\left(y, f(\mathbf{x})\right)\right] \\ &= \arg\min_{f:\mathcal{X}\to\mathbb{R}^g} \int L\left(y, f(\mathbf{x})\right) \mathrm{d}\mathbb{P}_{xy}. \end{aligned}$$

The resulting risk is called **Bayes risk**

$$\mathcal{R}_L^* = \inf_{f:\mathcal{X}\to\mathbb{R}^g} \mathcal{R}_L(f)$$

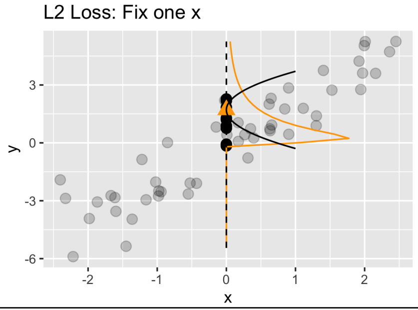
OPTIMAL POINT-WISE PREDICTIONS

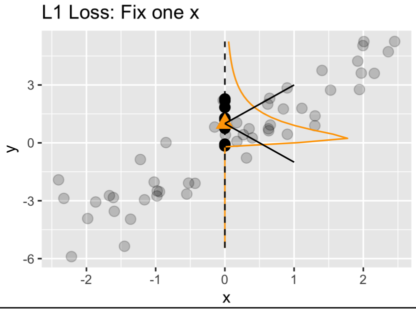
To derive the risk minimizer, observe that by the law of total expectation

$$\mathcal{R}_L(f) = \mathbb{E}_{xy}\left[L\left(y, f(\mathbf{x})\right)\right] = \mathbb{E}_x\left[\mathbb{E}_{y|x}\left[L\left(y, f(\mathbf{x})\right) \mid \mathbf{x} = \mathbf{x}\right]\right].$$

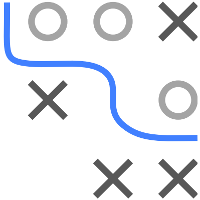
- We can choose $f(\mathbf{x})$ as we want (unrestricted hypothesis space, no assumed functional form)
- Hence, for a fixed value $\mathbf{x} \in \mathcal{X}$ we can select **any** value c we want to predict. So we construct the **point-wise optimizer**

$$f^*(\mathbf{x}) = \text{argmin}_c \mathbb{E}_{y|x}\left[L(y, c) \mid \mathbf{x} = \mathbf{x}\right] \quad \forall \mathbf{x} \in \mathcal{X}.$$





THEORETICAL AND EMPIRICAL RISK



The risk minimizer is mainly a theoretical tool:

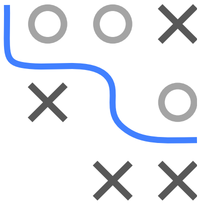
- In practice we need to restrict the hypothesis space $\mathcal{H}$ such that we can efficiently search over it.
- In practice we (usually) do not know $\mathbb{P}_{xy}$. Instead of $\mathcal{R}(f)$, we are optimizing the empirical risk

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f) = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right)$$

Note that according to the **law of large numbers** (LLN), the empirical risk converges to the true risk (but beware of overfitting!):

$$\bar{\mathcal{R}}_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right) \overset{n \to \infty}{\longrightarrow} \mathcal{R}(f).$$

ESTIMATION AND APPROXIMATION ERROR



Goal of learning: Train a model $\hat{f}$ for which the true risk $\mathcal{R}_L\left(\hat{f}\right)$ is close to the Bayes risk $\mathcal{R}_L^*$. In other words, we want the **Bayes regret**

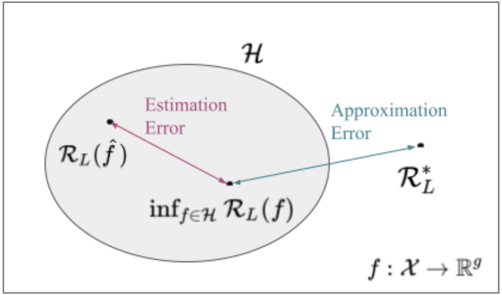
$$\mathcal{R}_L\left(\hat{f}\right) - \mathcal{R}_L^*$$

to be as low as possible.

The Bayes regret can be decomposed as follows:

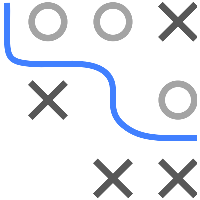
$$\mathcal{R}_L\left(\hat{f}\right) - \mathcal{R}_L^* \quad = \quad \underbrace{\left[\mathcal{R}_L\left(\hat{f}\right) - \inf_{f \in \mathcal{H}} \mathcal{R}_L(f)\right]}_{\text{estimation error}} + \underbrace{\left[\inf_{f \in \mathcal{H}} \mathcal{R}_L(f) - \mathcal{R}_L^*\right]}_{\text{approximation error}}$$

ESTIMATION AND APPROXIMATION ERROR / 2



- $\mathcal{R}_L\left(\hat{f}\right) - \inf_{f \in \mathcal{H}} \mathcal{R}(f)$ is the **estimation error**. We fit $\hat{f}$ via empirical risk minimization and (usually) use approximate optimization, so we usually do not find the optimal $f \in \mathcal{H}$.
- $\inf_{f \in \mathcal{H}} \mathcal{R}_L(f) - \mathcal{R}_L^*$ is the **approximation error**. We need to restrict to a hypothesis space $\mathcal{H}$ which might not even contain the Bayes optimal model f^*.

(UNIVERSALLY) CONSISTENT LEARNERS



Consistency is an asymptotic property of a learning algorithm, which ensures the algorithm returns **the correct model** when given **unlimited data**.

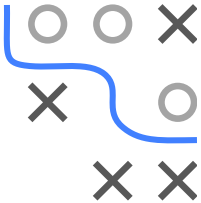
Let $\mathcal{I} : \mathbb{D} \to \mathcal{H}$ be a learning algorithm that takes a training set $\mathcal{D}_{\text{train}} \sim \mathbb{P}_{xy}$ of size n_{train} and estimates a model $\hat{f} : \mathcal{X} \to \mathbb{R}^g$.

The learning method $\mathcal{I}$ is said to be **consistent** w.r.t. a certain distribution $\mathbb{P}_{xy}$ if the risk of the estimated model $\hat{f}$ converges in probability ("$\xrightarrow{p}$") to the Bayes risk $\mathcal{R}^*$ when n_{train} goes to ∞:

$$\mathcal{R}\left(\mathcal{I}\left(\mathcal{D}_{\text{train}}\right)\right) \xrightarrow{p} \mathcal{R}_L^* \quad \text{for } n_{\text{train}} \to \infty.$$

(UNIVERSALLY) CONSISTENT LEARNERS / 2

Consistency is defined w.r.t. a particular distribution $\mathbb{P}_{xy}$. But since we usually do not know $\mathbb{P}_{xy}$, consistency does not offer much help to choose an algorithm for a particular task.

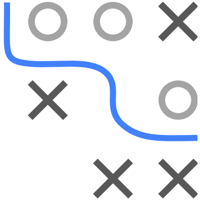


More interesting is the stronger concept of **universal consistency**: An algorithm is universally consistent if it is consistent for **any** distribution.

In Stone's famous consistency theorem from 1977, the universal consistency of a weighted average estimator as KNN was proven. Many other ML models have since then been proven to be universally consistent (SVMs, ANNs, etc.).

Note that universal consistency is obviously a desirable property - however, (universal) consistency does not tell us anything about convergence rates ...

OPTIMAL CONSTANT MODEL



While the risk minimizer gives us the (theoretical) optimal solution, the **optimal constant model** (also: featureless predictor) gives us a computable empirical lower baseline solution.

The constant model is the model $f(\mathbf{x}) = \boldsymbol{\theta}$ that optimizes the empirical risk $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$.